Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls

Panagiotis Papadopoulos
Brave Software

Peter Snyder
Brave Software

Dimitrios Athanasakis
Brave Software

Benjamin Livshits
Brave Software
Imperial College London

ABSTRACT

Funding the production of quality online content is a pressing problem for content producers. The most common funding method, online advertising, is rife with well-known performance and privacy harms, and an intractable subject-agent conflict: many users do not want to see advertisements, depriving the site of needed funding.

Because of these negative aspects of advertisement-based funding, *paywalls* are an increasingly popular alternative for websites. This shift to a “pay-for-access” web is one that has potentially huge implications for the web and society. Instead of a system where information (nominally) flows freely, paywalls create a web where high quality information is available to fewer and fewer people, leaving the rest of the web users with less information, that might be also less accurate and of lower quality. Despite the potential significance of a move from an “advertising-but-open” web to a “paywalled” web, we find this issue understudied.

This work addresses this gap in our understanding by measuring how widely paywalls have been adopted, what kinds of sites use paywalls, and the distribution of policies enforced by paywalls. A partial list of our findings include that (i) paywall use has increased, and at an increasing rate (2× more paywalls every 6 months), (ii) paywall adoption differs by country (e.g., 18.75% in US, 12.69% in Australia), (iii) paywall deployment significantly changes how users interact with the site (e.g., higher bounce rates, less incoming links), (iv) the median cost of an annual paywall access is 108 USD *per site*, and (v) paywalls are in general trivial to circumvent.

Finally, we present the design of a novel, automated system for detecting whether a site uses a paywall, through the combination of runtime browser instrumentation and repeated programmatic interactions with the site. We intend this classifier to augment future, longitudinal measurements of paywall use and behavior.

CCS CONCEPTS

• **Social and professional topics → Surveillance**; • **Information systems → Web applications**; • **Security and privacy →** *Economics of security and privacy.*

KEYWORDS

Paywalls, User privacy, Web Monetization, User Subscription

ACM Reference Format:
Panagiotis Papadopoulos, Peter Snyder, Dimitrios Athanasakis, and Benjamin Livshits. 2020. Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380217

1 INTRODUCTION

Publishers are increasingly moving away from ad-based models, because of the well-known failures [41] of ad-based internet funding models. The most common adopted alternative is for sites to deploy “paywalls”. “Paywalls” here are a broad term for monetization systems where visitors are charged subscription fees to access site content, sometimes after being able to sample a small amount of content for free. The upsides of paywall systems are well understood (i.e., they promise to enable the continued creation of high-quality content). Less understood are the risks and larger implications of an increasingly “walled” web. Possible risks include reducing societal access to news and information and the privacy harms of the increased user tracking needed to enforce paywalls.

This work aims to improve the understanding of the popularity, risks and benefits of paywalls online. To introduce the topic, we first (i) describe why the web is increasingly moving away from “open” models to “paywalled” models, (ii) outline why this transition is an important topic of study for the research community, and then (iii) present the structure of the rest of the paper.

1.1 The Move from Ads to Paywalls

Digital advertising is the current dominant monetization method for web publishers, and funds much of the web. Publishers sell advertisements along page content; middle parties buy these ad slots and fill them with images and content provided by clients and ad-agencies. This process is usually programmatic, based of user's personal (i.e., behavioral) data, and completed via real-time programmatic auctions [37, 42].

Web sites are increasingly unsatisfied from this ad-based funding system, for many reasons. First, the system is dominated by two parties, Google and Facebook, who jointly harvest more than 70% of global ad revenues [18, 46], reducing the publisher's “take” for ad placements through market power. Second, ad-based funding systems suffer from significant and increasing rates of fraud [8, 14, 15, 27, 64], depriving web sites of further funding. Third, behavioral advertising systems are increasingly incompatible with individual and

WWW '20, April 20–24, 2020, Taipei, Taiwan
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380217

legal privacy demands [26, 38, 40, 48, 59]. Last, users increasingly use ad blocking tools, for a variety of privacy, performance, and aesthetic reasons [34, 57], further depriving publishers of revenue. As a result, ad revenues have decreased in recent years. Both big and small publishers are coming up short on advertising revenue, even if they are long on visitors traffic. Accounts of publisher-loss under ad-based funding models contain figures as high as 95% [28].

The difficulties of ad-based funding systems have pushed publishers to alternative funding models, including donations [16, 63] or in-browser crypto-mining [39]. The most common alternative though is "paywalls", where users pay publishers directly to access the content they create [29] Figure 1 shows a representative example of a paywall system.

Paywalls so far have a mixed record as funding systems for publishers. Publishers with large, loyal audiences and high-quality content tend to be successful with this subscription strategy, with The New York Times [17], Wired [3], The Financial Times [9] and The Wall Street Journal [61] as prominent successful examples. The success of paywalls for smaller and more targeted sites (e.g., local news), or sites with less affluent audiences, is less clear.

It is important to note that the rapid growth of paywalls has drawn the attention of big tech companies like Google, Facebook and Apple, who have started building platforms to provide or support paywall services [25, 49, 51, 58], in an effort to claim their share of the market.

1.2 Understanding the State of Paywalls

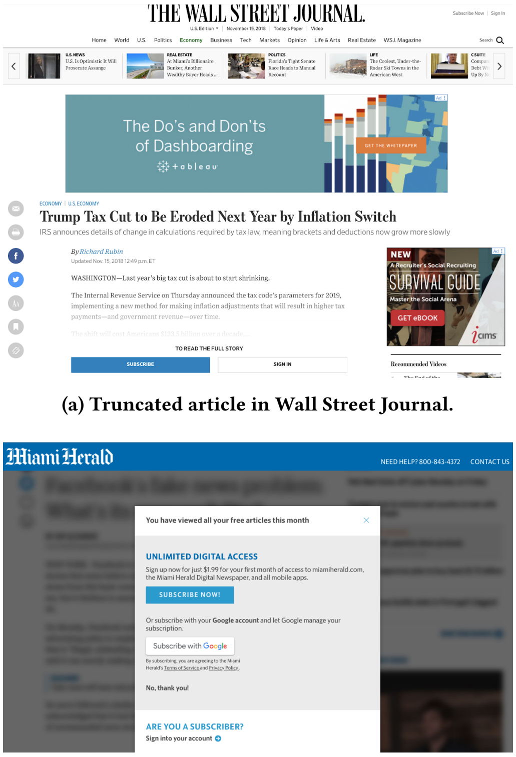
Creating a sustainable system to fund news and related content is an important goal, and paywalls seem to be a promising (partial) solution to the problem. However, this move from "open" to "walled" business strategies brings significant, understudied risks. For example, paywalls (implicitly or otherwise) may impose a "class system" on the web [11, 50], potentially driving information-seeking visitors who cannot afford to pay for subscriptions to badly-sourced, less-vetted, or even intentionally false (but free) new sources.

Despite the importance of the rise of paywalls to the web, it is surprising how little the topic has been studied by the research community. Important open questions include how popular paywall systems are, what policies paywalls impose, how users are tracked for paywall enforcement, and whether paywalls are effective at protecting premium content.

1.3 Contributions

In this work, we aim to improve the understanding of paywall systems through the first systematic study of paywalls on widely-used web sites. This work makes the following contributions to the understanding of paywall systems on the web:

(1) A **novel system for programmatically determining if a site is using a paywall**, though the combination of multiple crowd-sourced data sets and tools.
(2) A **case study of how a popular paywall library operates**, from how a publisher deploys it, how the paywall identifies users, to how the configured content access policy is enforced.
(3) A **large-scale measurement of paywall popularity**, including what kinds and what countries account for most paywall use, and how paywall use has changed over time. Example results include finding that paywall use has increased dramatically over time (2× more paywalls every 6 months) and that paywall adoption differs by country (e.g., 18.75% in US, 12.69% in Australia) and industry.
(4) An **in-depth, large scale analysis of deployed paywall policies**, including subscription costs, how paywall adoption impacts the hosting website, how robust paywalls are to evasion, the mechanisms paywalls use to prevent users from viewing protected content, and the privacy implications of paywalls.
(5) A **classifier for deterring whether a site is using a paywall** for use on sites not considered by crowd-sourced resources, for future long term, web scale measurements of paywall adoption and behavior.


THE WALL STREET JOURNAL.
Subscribe Now | Sign In
U.S. Edition | November 15, 2018 | Today's Paper | Video
Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine Search
The Do's and Don'ts of Dashboarding
tableau
GET THE WHITEPAPER
ECONOMY | U.S. ECONOMY
NEW
A Recruiter's Social Recruiting SURVIVAL GUIDE
Master the Social Arena
GET eBOOK
icims
Recommended Videos

Trump Tax Cut to Be Eroded Next Year by Inflation Switch

IRS announces details of change in calculations required by tax law, meaning brackets and deductions now grow more slowly

By Richard Rubin
Updated Nov. 15, 2018 12:49 p.m. ET

WASHINGTON—Last year's big tax cut is about to start shrinking.

The Internal Revenue Service on Thursday announced the tax code's parameters for 2019, implementing a new method for making inflation adjustments that will result in higher tax payments—and government revenue—over time.

The shift will cost Americans $133.5 billion over a decade,...

TO READ THE FULL STORY

SUBSCRIBE | SIGN IN

(a) Truncated article in Wall Street Journal.

You have viewed all your free articles this month

UNLIMITED DIGITAL ACCESS

Sign up now for just $1.99 for your first month of access to miamiherald.com, the Miami Herald Digital Newspaper, and all mobile apps.

SUBSCRIBE NOW!

Or subscribe with your **Google account** and let Google manage your subscription.

Subscribe with Google

By subscribing, you are agreeing to the Miami Herald's Terms of Service and Privacy Policy .

No, thank you!

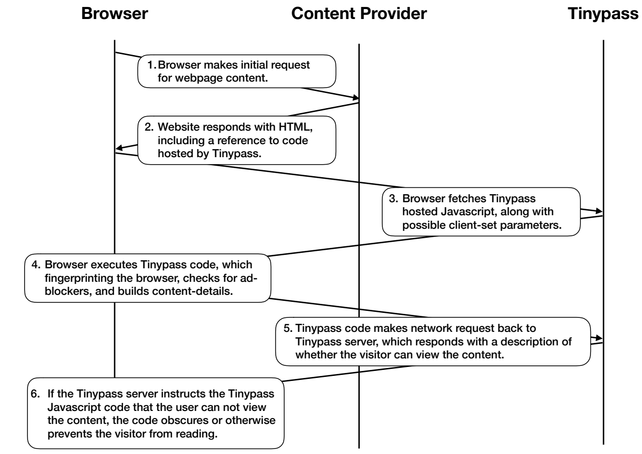
ARE YOU A SUBSCRIBER?
Sign into your account

(b) Obscured article in Miami Herald.

Figure 1: Examples of raised paywalls in major news sites. Paywalls may be enforced in different ways to deny access to articles to non-subscribed users.

2 BACKGROUND

Paywalls are an increasingly popular monetization strategy for web sites, as publishers attempt to become less dependent on advertising. Figure 1 shows a typical paywall, where a publisher is blocking access to content until the user pays a fee. To enforce access control, paywalls track the engagement of the user with the publisher content: i.e., how much time they spend on a web site, how many articles they have read, how many times a user has visited the website.



1. Browser makes initial request for webpage content.

2. Website responds with HTML, including a reference to code hosted by Tinypass.

3. Browser fetches Tinypass hosted Javascript, along with possible client-set parameters.

4. Browser executes Tinypass code, which fingerprinting the browser, checks for ad-blockers, and builds content-details.

5. Tinypass code makes network request back to Tinypass server, which responds with a description of whether the visitor can view the content.

6. If the Tinypass server instructs the Tinypass Javascript code that the user can not view the content, the code obscures or otherwise prevents the visitor from reading.

Figure 2: High level overview of the core functionality of a paywalled website powered by Tinypass.

2.1 Types of Paywalls

We group paywalls into two categories, based on how restrictive they are: (i) *hard paywalls*, where users cannot gain access the site without first purchasing a subscription (e.g., monthly or annual subscriptions) and (ii) *soft paywalls* that allow limited, free-of-charge viewing for a specific amount of time or number of visit (e.g., 5 free articles per month per user).

Hard Paywalls. Hard paywalls require subscriptions before visitors can access content (e.g., Financial Times requires a subscription before the user can read any article). Such a strategy runs the risk of deterring users and thereby diminishing the publisher's influence over all. As reported in the press [30], The Times experienced a 90% drop in traffic after introducing a hard paywall.

Soft or Metered Paywalls. Soft (or metered) paywalls limit the number of articles a viewer can read before requiring a paid subscription. Soft paywalls use the free articles as a strategy to entice users to subscribe. Soft paywalls require some method (often a JavaScript snippet on the user-side) for measuring either the number of articles a user has accessed, or the time a user spends in browsing the website's articles.

As with hard paywalls, a publisher's web traffic can also be affected by the installation of soft paywalls (e.g., traffic to the New York Times declined by 5% to 15% one month after the installation of its soft paywall [43, 52]). Overall, though, fewer users are discouraged by soft paywalls. Prior studies [21] have found, on average, retention rates for publishers with soft paywalls reaching 58.5%, compared to only 15–20% for publishers with hard paywall.

3 PAYWALL CASE STUDY

This section provides a detailed case study of a popular third-party paywall system. We provide this case study (i) to introduce the reader to how paywalls work, and (ii) to document the kinds of privacy-affecting behaviors paywalls often rely on to impose their policies. We select Piano's Tinypass paywall-as-a-service product [44] for our case study for several reasons. First, it is one of the most popular third-party paywall providers (Tinypass owns 38.2% of the market, as measured in Figure 11), so understanding how this system works provides a good understanding of the kinds of paywall code users are likely to experience. And second, Tinypass can be deployed as a configurable, paywall-as-a-service, allowing publishers (blogs, news sites, magazines, etc.) to impose a variety of paywall policies, both hard and soft.

```
var _getFingerprint = function () {
    if (fingerprint) {
        return fingerprint;
    }
    var fingerprint_raw = _getLocality();
    fingerprint_raw += _getBrowserPlugin();
    fingerprint_raw += _getInstalledFonts();
    fingerprint_raw += _getScreen();
    fingerprint_raw += _getUserAgent();
    fingerprint_raw += _getBrowserObjects();
    fingerprint = murmurhash3.x64hash128(fingerprint_raw)
        ;
    util.debug("Current browser fingerprint is: " +
        fingerprint);
    return fingerprint;
};
```

Listing 1: Excerpt of Tinypass's fingerprinting JavaScript.

3.1 Tinypass: The protocol

At some point prior to the user's visit, a site owner creates an account at Tinypass, where they describe the subscription policies they wish to enforce. Tinypass generates the keys and identifiers used to enforce the paywall and track visitors. Once a site owner installs Tinypass on their site, the paywall works in the following six stages, with numbers corresponding to Figure 2:

Step one. The user's browser makes a request to a website where the site owner has installed Tinypass.

Step two. The website responds with the HTML of their page, including a reference to the Tinypass JavaScript library, hosted on Tinypass's servers. The content provider's response may also include optional, customized parameters that allow Tinypass to integrate with other services, like Facebook and Google Analytics. At the time of this writing, Tinypass's code is hosted at https://code.tinypass.com/tinypass.js.

Step three. The referenced JavaScript causes the browser to request code from Tinypass's server, which responds with a bootstrapping system, providing basic routines for fetching the main implementation code, helper libraries, and utilities for rate limiting and fingerprinting. Depending on the particular deployment, minified versions of this code also includes common utilities like CommonJS-style dependency tools or cryptography libraries.

Step four. The browser executes the complete Tinypass library, and the full (post-bootstrap) Tinypass library performs a number of privacy-relevant checks. First, Tinypass attempts to determine if a site visitor is actually an automated browser (e.g., Puppeteer, WebDriver client). Tinypass attempts to determine if the user has an ad-blocker installed. Interestingly, Tinypass not only detects if the user currently has an ad-blocker installed, but also if the visitor has changed their ad-blocker usage (e.g., the user had an ad-blocker installed on a previous visit but no longer does, or vice versa).

Tinypass then generates a user fingerprint, implemented with the code hosted at https://cdn.tinypass.com/api/libs/fingerprint.js. The Tinypass fingerprinting library (shown in part in Listing 1)

```
...
    "trackingId": "{jcx}H4sIAAAAAAAAAI2QW2vCQBCF_8s...",
    "splitTests": [],
    "currentMeterName": "DefaultMeter",
    "activeMeters": [
        {
            "meterName": "DefaultMeter",
            "views": 0,
            "viewsLeft": 4,
            "maxViews": 4,
            "totalViews": 0
        }
    ],
    ...
```

Listing 2: Excerpt of returned Tinypass end point data (meter is Tinypass's terminology for a counter describing how much more non-paywalled content a user can view).

hashes together a number of commonly known semi-unique identifiers (installed plugins, preferred language, installed fonts, screen position, user agent, etc.) to build a unique identifier, hashed together using the MurmurHash3 hash algorithm [2]). The result is an identifier that is consistent across cookie-clears, and so can re-identify users attempting some evasion techniques. Tinypass also reads, if available, a first-party cookie the library also uses to identify users. When available, this cookie is used in place of the above fingerprint, to track how much content the user has visited.

Step five. Next, the Tinypass library gathers the above information, combines it with information about the page, derived fingerprinting values, the date, and other similar data, and POSTs them to a Tinypass endpoint [1], which records information about the page view. The server then returns a JSON string describing a variety of information about the page view, and excerpt of which is presented in Listing 2. This JSON string includes a wide variety of both user-facing and program-effecting values, including how many more pages the user is able to visit before the paywall is triggered, possibly new identifiers to rotate on the browsing session, whether the user has logged in and is known to Tinypass (e.g. the user logged in on a different domain owned by the same publisher).

Step six. Finally, the Tinypass code running on the browser enforces the described paywall policy. The code, client-side, uses the response data to decide how to respond to the page view, possibly by obscuring page content or presenting a subscription offer dialog (by default, Tinypass offers pre-made-but-configurable modal and "inline" dialogues the website can check from). In the pages we observed, Tinypass only enforced subscription requirements (i.e., preventing users from viewing content) after the above check was completed. A side effect of this implementation decisions is that Tinypass's restrictions can be circumvented by simply blocking the Tinypass library (see Section 5).

4 CURRENT PAYWALL DEPLOYMENTS

In this section, we present a large-scale measurement of paywall deployments on the web. The measurements presented give a broad assessment of how often paywalls are used (by country and by industry). We then present a variety of measurements of how deployed paywalls operate, including the access policies they enforce, their enforcement mechanisms, and how robust these paywalls are to circumvention. The section begins with a description of how we gathered there relevant datasets for measurements, and then proceeds in the above described order.

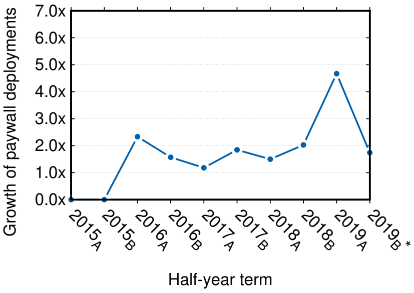


Figure 3: Growth of paywall deployments per 6 months. Note that the y-axis depicts the growth-rate, and not absolute numbers.

4.1 Dataset

To conduct the measurements described in this section, we built an oracle to determine whether a web site uses a paywall. While seemingly a simple question, the diversity of paywall libraries, enforcement mechanisms, access policies and varying verbiage makes this a difficult question to answer without significant human intervention. To solve this problem, we draw on two existing crowd-sourced datasets:

A) Extensions. First, we extract rules from several popular browser extensions [1, 13, 24, 35] designed to help users circumvent paywalls. By examining the source code of these extensions, we are able to (directly or indirectly) identify 147 websites that the tools' authors and maintainers label as using paywalls.

B) Filter lists. Second, we use a popular, crowd-maintained filter list that identifies third-party paywall libraries [7] so that they can be blocked with common filter-list consuming tools (e.g., AdBlock Plus, uBlock Origin). This list includes filter rules for blocking resources related to a variety of internet "annoyances"; we extract the subset of the list specifically targeting paywalls. This gives us a list of 43 third-party paywall libraries. We query for each entry of this paywall libraries list in two existing, current web crawl archives (i.e., HTTPArchive [22] and PublicWWW [47]). We found 1,563 sites using one of these paywall libraries and we labeled them as "paywalled".

We combine the above two approaches (i.e., paywalled domains labeled by browser extensions and paywalled sites including third-party paywall libraries) to identify 1,710 unique paywall-using domains, from 61 countries. This dataset, summarized in Figure 4,

Data	Volume
Paywalled websites from bypassing extensions	147
Third-party paywall libraries	43
Unique paywalled sites	1,710
Countries the paywalled sites originate from	61

Figure 4: Summary of our crowdsourced dataset labeling which websites use paywalls.

[1] https://experience.tinypass.com/xbuilder/experience/execute?aid=*

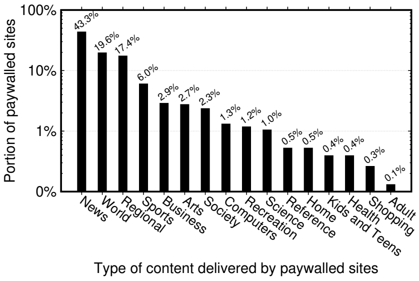


Figure 5: Type of industry or type of content, paywalled websites deliver.

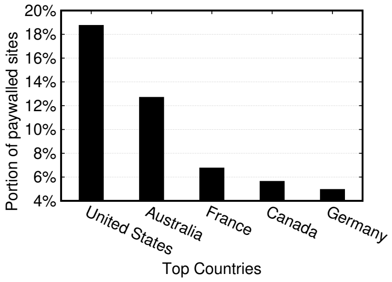


Figure 6: Portion of news sites using paywalls per country. Paywall adoption reaches 18.75% in US and 12.69% in Australia.

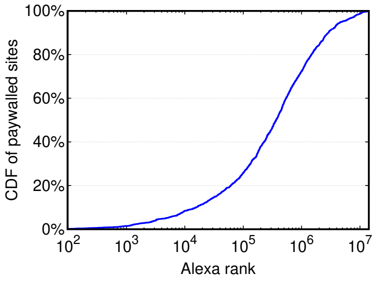


Figure 7: Distribution of the popularity of each paywalled website in our dataset based on the Alexa ranking.

comprises our complete dataset of paywalled sites used as an oracle in this section, and we provide open-sourced[2].

4.2 Paywall Popularity

We start by measuring how popular paywalls are, across several dimensions. We use the domains identified in Section 4 as the set of paywalled sites, and all other sites on the web as not paywalled.

4.2.1 Increase in Paywall Use. We first measure whether paywall use has increased over time. We find sites in our dataset started using paywalls in 2015, with overall paywall use strictly increasing since. Paywall use has increased at a rate between 120% and 230% every six months since 2015 until recently. In the first six months of 2019, paywall use quadrupled, and has grown by a further *180%* during the first two months in the second half of 2019. These measures are summarized in Figure 3.

We measure paywall growth over time by applying our paywall oracle (described in Section 4.1) to archived versions of the same sites in the Wayback Machine web archive [22]. We use these archived versions of each site to approximate date each site adopted a paywall. The precise methodology is as follows:

(1) We build the set of paywall library related URLs and domains using the technique described in Section 4.1.
(2) We fetch the most recent archive of each paywalled website in our dataset from the Wayback machine and check whether that historical version is using a paywall.
(3) If we observe the site using a paywall, we fetch the next-most-recent version of the site from the Wayback machine (e.g., we move back one recording in time) and re-check.
(4) We continue this process until we encounter a version of the web site that no longer is using a paywall. Once we encounter a non-paywalled version of the site, we note the date that version of the site and record it as when the site began using a paywall.

Limitations. We note two limitations of the above approach, and why we do not believe they significantly impact our findings. It is possible that earlier versions of sites used different types and providers of paywalls than current sites, and so our paywall detection oracle may be missing historical paywall use. While possible, we do not think this limitation significantly impacts the results for two reasons: (i) prior research [60] has found that filter lists (like the ones we use for paywall library detection) rarely delete rules, and so that paywall-targeting filter lists would identify both current and historical paywalls. What is more, (ii) we manually evaluated a random sample of commits from the git history of the paywall-targeting portion of the filter list and we found no rule deletions. This gives us further confidence, though not certainty, that filter rules that would identify paywalls on previous versions of paywalled sites have not been removed.

The second possible limitation is that our approach might miss sites that used to have paywalls, but no longer do. We believe such cases to be rare. We observed no instances of sites using paywalls, removing the paywalls, and then re-establishing it. This suggests (though does not prove) that sites do not commonly abandon paywall strategies once they have adopted them.

4.2.2 Paywall Use by Site Type. We measure what types of content paywalled sites provide. We find that most (80.3%) paywalled sites provide some form of news content, whether targeted at the local, regional, or world-level. Figure 5 provides summary of this measurement. For this measurement, we use the sites identified as using paywalls from Section 4.1 with information available from the Alexa Top Sites service. The Alexa Top Sites classifies domains into one of 17 different classes (i.e., news, sports, business, arts, society). Three categories describe news content, though at different levels of focus (e.g., "World", "Regional" or, generically, "News"). We group these together for our measurements, since they are thematically very similar. The remaining 14 categories account for just 19.7% of paywalled sites.

4.2.3 Paywall Use by Country. Next, we measure which countries have the highest rates of paywall use. Because news sites account for most paywall use, we focus this measurement on news sites. We find that US news sites have been the quickest to move to paywalls, followed by Australia, France, Canada and Germany. Figure 6 summarizes our findings. Since our oracle does not identify all websites with paywalls, Figure 6 presents only the lower bound of the existing paywalled sites.

We measure rates of paywall use by country by first retrieving the Alexa the Top 10,000 websites per country. We filter the list and remove all non-news sites. Then, we calculate the percentage of paywall-using news sites, as a fraction of all news sites, per

[2] https://gist.github.com/panpap/68af1c99b49366dfce4044a354f6e1b8

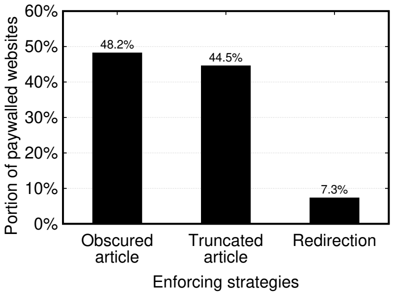


Figure 8: Popularity of the different paywall enforcing policies. Most of the publishers prefer to obfuscate (48.2%) or truncate (44.5%) the article the user has not yet access to.

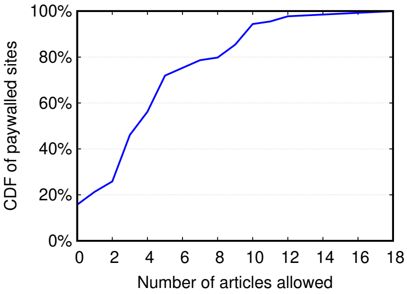


Figure 9: Distribution of the number of free articles allowed per user. The median paywalled website allows 3.5 articles, the median soft-paywalled website allows 4 articles and all hard-paywalled do not allow *any* free article.

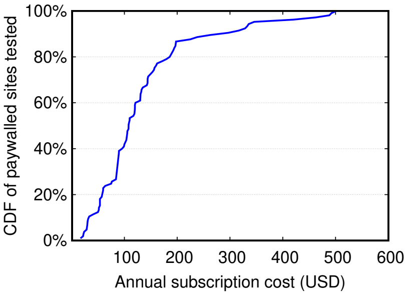


Figure 10: Cumulative distribution of the subscription cost per website for a 12-month content access.

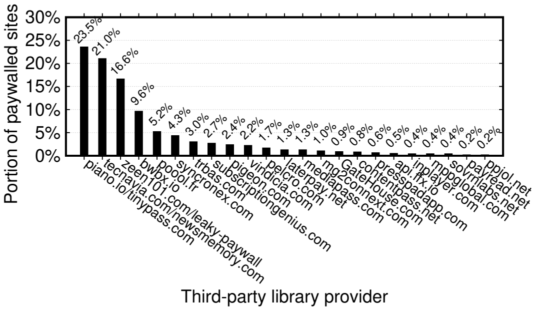


Figure 11: Popularity of third-party paywall libraries in our dataset. A small number of paywall implementations account for the majority of third-party paywall deployments.

country. We find that 18.75% of US news sites use paywalls, 12.69% of Australian new sites, and less than 7% in all other countries.

4.2.4 Paywall Use by Popularity. Next, we measure whether there is any clear relationship between paywall deployment and the popularity of a website. We did not observe any such relationship. We anticipate that most paywalled sites would be popular (as measured by Alexa Top Sites), as a successful paywall would require a significant number of subscribers, which in turn would require a significant amount of baseline visitors. Instead, we find that only 8.54% of paywall-using sites are among the 10,000 most popular sites on the web. The median paywall-using site is ranked 365,316. The full distribution of the popularity of paywall using sites is presented in Figure 7.

4.3 Paywall Libraries

A significant number of sites rely on third-parties for their paywall implementations. These third-parties sell “paywall-as-a-service” products, where publishers pay fee to have the third-party manage and enforce the paywall on the publisher’s site. We observe that a small number of paywall providers account for the vast majority of paywall deployments, with Piano and Tecnavia being the most popular paywall providers (23.5% and 21.0% market share, respectively). The full distribution of third-party paywall market share is depicted in Figure 11.

This consolidation of paywall implementation and enforcement is significant, for a variety of reasons. First, market consolidation may effect the amount of income content-makers can receive for their content (popular third-party paywall providers receive 10-15% of each sold subscription). Second, provider consolidation may make large scale circumvention easier, as circumventors need to target a smaller number of systems (see Section 5). Third, a small number of paywall providers tracking users across a large number of websites has clear privacy implications (see Section 4.6).

We measure the popularity and consolidation of third-party paywall libraries by crawling each paywalled site in our dataset and observing which resources from known paywall providers were fetched. We find that at least 25% of paywalled websites outsource their paywall functionality to third-parties. The distribution of third party paywall use follows a rough power-law distribution.

4.4 Paywall Polices

Next, We measure the distribution of policies enforced by paywalls. We find that paywalls vary widely by type, enforcement mechanism, and how much, if any, content visitors can view before needing to pay. For these measurements, we randomly sample 115 paywall-using websites from our dataset for manual evaluation.

4.4.1 The different types of Paywalls. First, we observe that 66.7% of paywalls are “soft” (i.e., allow some free content access), 15.7% are “hard” (i.e., allow no free access), with the remaining 16.6% paywalled sites using a “hybrid” strategy (i.e., some content is free, some requires payment, based on the author/time of publication/topic, etc.). Some “hybrid” publishers use machine learning or other dynamic approaches to determine per-user whether an article should be locked or not [19, 45, 56].

4.4.2 Enforcement Mechanism. We also measure the distribution of paywall enforcement techniques. Despite the heterogeneity of

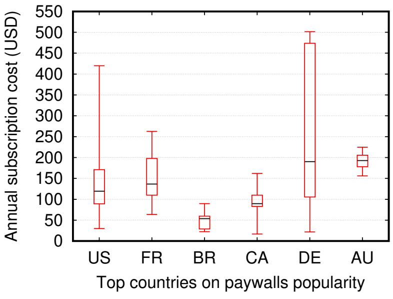


Figure 12: Min, 15th percentile, median, 85th percentile, and max annual subscription costs for paywalls, by country.

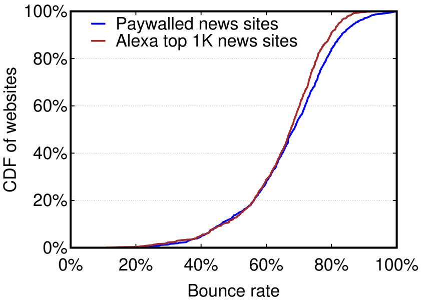


Figure 13: Distribution of the bounce rate per website. The median paywalled site has slightly higher bounce rate (68.4%) contrary to the non-paywalled (67.5%).

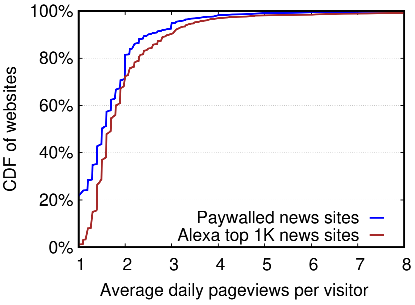


Figure 14: Distribution of the daily page views per visitor per news site. In median values, an average visitor browses on a daily basis 13.42% less pages on a paywalled news site.

the paywall implementations, we see only three approaches used to enforce paywalls: (i) truncating article text, (ii) obfuscating the article with popups, or (iii) redirecting users to a subscription page. We measure the popularity of each of the above approaches in our manually evaluated set; Figure 8 presents the results. The largest percentage (48.2%) of the websites obfuscate or truncate (44.5%) the article the user has not yet access to. Only a few (7.3%) redirect the user to a login/subscribe page.

4.4.3 Allowed Free Content. We also measure the distribution of how much content users can view before triggering a (hard or soft) paywall. For the 15.7% of sites that use a hard-paywall strategy, visitors cannot view *any* articles for free. For soft paywalls, this number varies by publisher. Figure 9 plots the distribution of the free articles users could consume before hitting a paywall in the websites we tested. Overall, the median paywalled website allows 3.5 articles. All hard paywalled websites do not allow *any* access to articles, when the median soft-paywalled website allows 4 articles to be read for free. A significant number of soft paywalls (30%) that allow 2 or fewer articles to be read before triggering enforcement.

4.4.4 Paywalls Cost. Next, we measure the distribution of paywall subscription costs. We find that most paywall subscriptions are monthly, that the median annual cost for paywall access is 108 USD, and that subscription costs seem to be highest in Germany. All of these measurements were conducted through a manual evaluation of 105 paywalled sites. We sampled 20 sites from each of the top 6 paywall using countries. For 12 sites, we were not able to access the site or determine the subscription costs

We first measure the distribution of subscription options for users. 82.86% of paywall sites provide a monthly subscription option and 35.23% of sites provide an annual one. Hence, 64.76% of the paywalled sites provide *only* a monthly subscription option and 17.14% *only* an annual one. Next, we measure the distribution of purchasing an annual subscription to a site's content. The median observed annual subscription cost is 108 USD. 22% of sites charge less than 60 USD, 21% of sites charge more than 180 USD. Figure 10 presents the full distribution of annual subscription costs. We note that the subscription rates we observe are lower than those estimated by previous work (around 189 USD on average) [55], possibly reflecting a general decrease in costs. We measure the distribution of annual costs by manually noting the annual subscription cost in the local currency. For sites that do not offer an annual subscription, we sum the cost of twelve monthly subscriptions. We then convert all costs to USD for comparison purposes.

Finally, we measure how subscription costs differ by country. As depicted in Figure 12, we plot the min, the 15th percentile the median, the 85th percentile and the max of the annual subscription cost across the different countries. The median prices of subscriptions in Australia and Germany are highest (193 and 190 USD, respectively). Subscription costs vary widely by site, too. In Germany and the United States, for example, the most expensive paywalls cost 2.63× and 3.51× more than the median rate, respectively.

4.5 How Paywalls Affect Site Use

Paywalls affect how users interact with the site. Recent studies [23], by monitoring the pageviews of 36 news sites before and after paywall deployment, report that they lose nearly 30% of their daily traffic and a loss of pageviews, ranging from a 10% to 55%. In this section, we measure differences between how sites interact with paywalled and non-paywalled sites. We find that users view less pages on paywalled sites, stay for shorter periods of time and link to pages less. Interestingly, we did not see a significantly difference to the bounce rate between paywalled and non-paywalled sites [3].

4.5.1 Bounce Rate. We find that paywalled new sites have slightly higher bounce rates [4] than non-paywalled news sites. The distributions of bounce rates is depicted in Figure 13. The median paywalled news site has slightly higher bounce rate (68.4%) contrary to the median non-paywalled (67.5%). However, we see that for 30% of the websites in the two categories the difference is 2-7% higher.

[3] We do not address the issue of causation; its possible, for example, that the types of site likely to use paywalls have lower *dwell times* already, so that the use of a paywall is a more a result of lower dwell time than the cause. We leave disentangling cause and effect for future work.

[4] The percentage of visitors who visit a site and then leave, rather than continuing to view other pages within the same site.

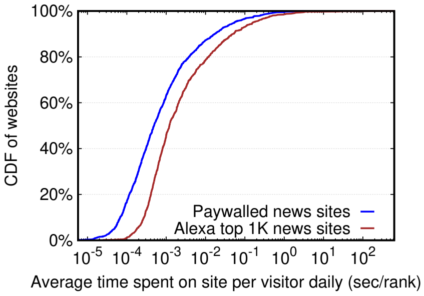


Figure 15: Distribution of the average time a visitor spends daily per news site. In median values visitors tend to spend daily 2.46× more time per site rank on non-paywalled websites.

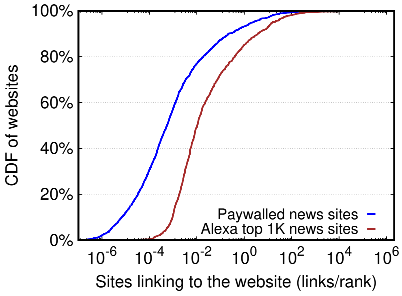


Figure 16: Distribution of the incoming site links per news site. Paywalled sites get significantly (18.9×) less site links per rank in median values.

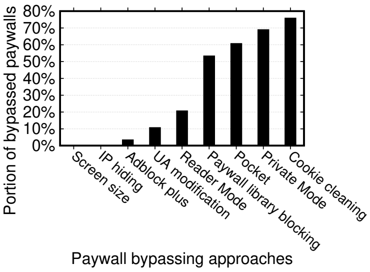


Figure 17: Success rate of the different paywall bypassing approaches. Clearing the cookie jar alone can bypass 75% of the paywalls.

To compare bounce rates, we used the Alexa Top Sites data, and compared the bounce rates for the paywalled news sites in our data set with the Alexa top 1K news sites.

4.5.2 Daily Page Views. Next, we measure the number of pages the average visitor performs daily on the websites and we compare how this changes for the paywalled and non-paywalled news sites. In Figure 14, we plot the cumulative distribution of these page views per website in our dataset. Users visit on average 13.42% less pages on paywalled news sites than non-paywalled new sites.

4.5.3 Average Time spent on Site. Figure 15 compares the distribution of the median time users spend on paywalled and non-paywalled websites, normalized by popularity (based on its Alexa rank). We find that visitors spend daily 2.46× more time per on non-paywalled news sites.

4.5.4 Content Popularity and Link Rate. Finally, we measure the impact of paywalls on how often sites link to the paywalled sites. Since site linking may be affected by the popularity of the website, in Figure 16, we plot the cumulative distribution of the number of site links (or backlinks) per news site normalized by its Alexa rank. We observed paywalled sites being linked to significantly less (18.9×) often than non-paywalled sites.

4.6 Paywalls and Privacy

Most behavioral advertising systems require users to pay for content with their privacy; users are tracked in behavioral advertising systems, and can view "free" content. Paywalls have the possibility of changing this system. Since users are directly paying for content, one might hope users would no longer face the privacy harms associated with behavioral advertising systems. Unfortunately, we see that this is not the case. People *do not* generally receive a tracker free version of site content when paying for subscriptions. Instead, paywall systems seem to serve as an *additional* monetization mechanism on top of existing, privacy harming, ad systems.

We measure whether paying for paywall access improves user privacy (i.e., removes the need for sites to try and monetize through tracking) by purchasing subscriptions to 10 randomly selected paywalled news sites. Our goal is to examine the types of network requests issued before and after paying for the subscription. We create two scenarios, the vanilla (non-subscribed) user, and the premium (subscribed) user. For each selected site, we create an account and purchased a subscription before the starting the measurement. We also select 5 child pages on each site for evaluation. Then, we enable the popular Disconnect plugin [5] in monitoring and no-blocking mode, and browse each selected child page on each site under each of the two personas, in the same order, and observe the issued network requests. Figure 18 presents the average number of ad- and tracking- related requests encountered in each scenario. In the vast majority of cases, there is no significant difference in terms of ad- or tracking- related web requests.

5 PAYWALL CIRCUMVENTION

Paywalls must be robust to circumvention if they are going to be a successful monetization scheme for websites. If paywalls can be easily avoided, then content producers will wind up in the same situation they are in with ads and ad-blockers; declining revenues as circumvention tools become more popular. We find that *all observed paywalls are trivial to circumvent.*

We evaluate how robust paywalls are to circumvention in two steps: (i) we categorize the approaches of several popular paywall

	Vanilla User		Premium User	
News site	**Ads**	**Tracking**	**Ads**	**Tracking**
heraldsun.com.au	171	13	169	9
miamiherald.com	123	12	112	11
wsj.com	63	4	61	4
kansascity.com	61	9	56	6
ft.com	20	0	11	0
salon.com	138	5	0	1
japantimes.co.jp	109	12	98	8
leparisien.fr	125	10	81	4
independent.co.uk	11	6	10	6
spectator.co.uk	18	2	14	2

Figure 18: Requests captured for vanilla and premium user. User continues receiving the same amount of trackers and ads in the content she receives even if she has paid for it.

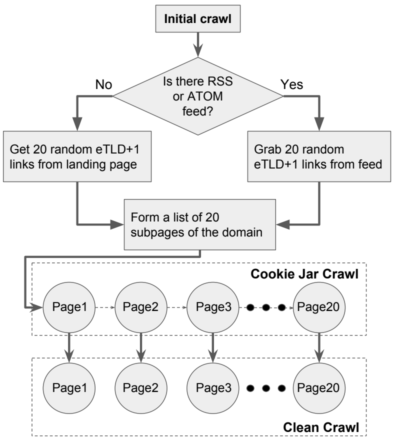

Initial crawl
Is there RSS or ATOM feed?
No
Yes
Get 20 random eTLD+1 links from landing page
Grab 20 random eTLD+1 links from feed
Form a list of 20 subpages of the domain
Cookie Jar Crawl
Page1
Page2
Page3
Page20
Page1
Page2
Page3
Page20
Clean Crawl

Figure 19: Data collection steps of our paywall detector's crawling component.

circumvention strategies, and (ii) we test each strategy on 32 paywalled news sites, we randomly select from our dataset. This subset comprises 28 soft and 4 hard paywalls on popular websites like Wired, Bloomberg, Spectator, Washington Post, Irish Times, Medium, Build, Japan Times, Statesman, and Le Parisien.

5.1 Evasion Evaluated

We test the robustness of each paywall system by using Chrome version 71. For each evaluated site, (a) we browse different pages till we trigger the paywall, and then (b) we test a variety of bypassing approaches to circumvent the paywall and get access to the "protected" article. Figure 17 lists the evaluated paywall-circumvention strategies, which includes pre-packaged tools, fingerprint evasion techniques, and third-party services. Specifically, we consider:

(1) changing the screen size dimensions
(2) hiding the user's IP address
(3) changing the user agent string
(4) using an ad blocker extension
(5) enabling "Reader Mode"
(6) using the Pocket web service[5]
(7) enabling Incognito/Private Mode
(8) emptying the cookie jar
(9) blocking HTTP requests for popular paywall libraries

Overall, we are able to bypass *all of the soft paywalls and none of the hard paywalls*. Hard paywalls perform their enforcement server-side, when the soft paywalls perform their policy enforcement client-side, and thus their access control is circumventable.

5.2 Evasion Approaches Analyzed

Many of the evaluated evasion approaches are rarely successful. For example, changing the screen size or the IP address of the user rarely circumvents a soft paywall (4% effectiveness). A moderate number of soft paywalls (12%) is flummoxed by modifying the browser's user agent string. The majority (75%) of soft paywalls is bypassable by resetting the cookie jar (in some cases erasing the first-party cookie only is insufficient, since it is automatically re-spawned by fingerprinting JavaScript code, as seen in Section 3). As a result, switching into browsers' "private browsing" modes is also sufficient to bypass most paywalls. Some paywalled sites refuse to render content in "reader modes" or "private browsing" modes, either first party (e.g., reader modes shipped with Safari and Firefox) or third-party (e.g., services like Pocket [62]). Such detection schemes are uncommon though; switching into reader-mode, for example, circumvents paywall enforcement in 60% of the cases. Ad-blocking extensions, in their default configurations, have little-to-no effect on paywalls. However, by using the list of known paywall libraries from Section 4 and by blocking requests to these domains we are able to bypass 48% of the paywalls without breaking the website's main functionality.

Metric	Value
Precision	77%
Recall	77%
F-Measure	75%
AUROC	0.74

Figure 20: Weighted average of the performance of our RF classifier, after k=5 cross-fold validation.

Third-parties like Google Search, Twitter, Reddit and Facebook, can also be used to gain access to some paywalled articles. Some paywalls give visitors from these large third-party systems unfettered access to their content, in pay-for-promotion initiatives. By spoofing the referrer field of the HTTP GET requests, some paywalls are vulnerable to exploiting a controversial policy [4] where publishers (for promotion purposes) allow access to articles when the visitor comes from one of these platforms (by clicking on a tweet, a post, a Google search result etc.) [6]. These mechanisms can provide access to hard paywalled articles. As a result, some publishers (e.g., Wall Street Journal) have stopped allowing such special access through their paywalls [31].

6 PAYWALL DETECTION

This section presents the design and evaluation of a ML-based detection system whose goal is to determine whether a site uses a paywall. Our paywall detector consists of two components: (i) a crawling component that visits a subset of pages on a site, records information about each page's execution, and extracts some ML features; and (ii) a classifier, that uses the extracted features to predict if the site uses a paywall.

We present this classifier as a partial solution for the problem of measuring changes in the adoption and behavior of paywalls over time. We propose this ML approach as a *complement* to the crowd-sourced approach described in Section 4.1. The classifier can be used to automatically gauge paywall prevalence. This automated approach can help identify and quantify paywalls that have not been identified by crowd-sourced lists, such as paywalls deployed by unpopular or region-specific sites.

Before describing the classifier in detail, we note two things. First, the classifier is designed to help detect broad, web-scale trends in paywall use and behavior, not to detect at real time paywall use on any single specific site. Second, an important finding of this classifier is that there is far greater diversity in paywall behavior and implementation logic than we expected at the start of the effort. We expect this to be a useful starting point for future studies.

[5] or similar "reader" services like "JustRead" [53] and "Outline" [36]

6.1 Crawling Methodology

The data collection step of our paywall detector, depicted in Figure 19, begins with three crawls of the target website: (i) the *initial crawl* that collects a list of child pages on the website, (ii) the *cookie jar crawl*, where each child page is crawled sequentially in the same browsing session and (iii) the *clean crawl*, where each child page is crawled with a fresh browsing session (i.e., a "clean" cookie jar).

This strategy replicates viewing patterns that might cause a paywall to be triggered, and then attempts to detect the paywall's presence by looking for page content that was visible on previous visits, but is no longer visible. For each page crawled, the crawler records the final state of the DOM, which DOM elements are visible, which page elements are positioned to obscure others (e.g., modal dialogs), and other page execution data only available at runtime.

6.2 Feature Extraction

We selected features that target both immediately triggering paywalls and paywalls that trigger after viewing multiple pages. These features aim to capture an intuition about how paywalls behave, and can fall into three rough categories: textual features, structural features, and visual features.

Text features. These features consider the text of the page, targeting text and idioms associated with paywalls. The crawler looks for the phrases "subscribe", "sign up" and "remaining" (translated into 87 languages) in (i) the "readermode" subset of the page, (ii) any overlay or popup elements (e.g., elements that have, or are children of elements that have, z-index values greater than zero), and (iii) elsewhere in the page. These three checks are performed both in "cookie jar" and the "clean crawl" recordings of each page.

Several text features use a "readermode" version of page, the subsection of the document identified as the page's "main content", or the content stripped of page "boilerplate" elements (e.g., advertisements, navigation elements, decorative images). While there are many different "readermode" identification strategies [20], in this work we use Mozilla's *Readability.js* [32] implementation, because of its popularity and ease of use. We expect using other "readermode" strategies would work roughly as well.

Structural features. These features target page structure (i.e., HTML), independent of specific page text or presentation. Structural features include whether the website has a RSS or ATOM feed, changes in the number of text nodes present in the page between its "cookie jar" and "clean crawl" versions, how many measured pages contain a "readermode" subset, and the average and maximum difference in the amount of text in the document in "readermode", between "cookie jar" and "clean crawl" measurements.

Visual features. These features focus on visual aspects of measured pages, and how those visual aspects change between the "cookie jar" and "clean crawl" measurements for each child page. The detector measures how many text nodes are obscured and the average and maximum change in obscured text nodes between the two measurements for each page. Additional display features are the number, and change in, text nodes in the browser viewport, and number of text nodes (regardless of text content) appearing in overlay (i.e., *z*-index great than zero) page elements. These features identify paywalls that prevent users from reading page content through popups or similar methods.

6.3 Classifier Accuracy

Our paywall detector uses a *random forest* classifier, specifically the *RandomForestClassifier* implementation provided by the popular SciKit-Learn [54] python library. Classification parameters were selected through 5-fold evaluation using the entirety of the aforementioned extracted features. As a ground truth, we use a subset of the paywalled sites the oracle identified (Section 4.1). To assess the accuracy of the classifier we use a different subset of the oracle's data and a set of non-paywalled websites we manually generate. The paywall detector achieves an average precision of 77%, recall of 77% and an area under the receiver operating characteristics (AUROC) of 0.74. These results are encouraging and suggest that our approach can be used to gauge paywall prevalence on the web. They also indicate that paywalls vary in behavior more than we anticipated, and that more complex features may be needed to further improve accuracy.

7 RELATED WORK

In [12], authors perform an empirical study of the pay models (freemium and paywall models) in European news. In particular, they manually analyzed a small dataset of 171 of the most important news organizations in France, Poland, Germany, Italy, Finland, and UK. Their results show that 66% percent of the newspapers operate a pay model and that the average price for a monthly subscription is 13.64 Euros when prices in general range from 2.10 to 54.27 Euros/month. In our measurements, 3 years after, the average monthly subscription cost 10.93 Euros, when specifically in Germany it is 20.48 (was 19.75) Euros and in France it is 12.54 (was 13.97) Euros.

In [33], authors explore the content that news publishers consider worthy of placing behind a paywall. They analyze 614 articles from the leading Australasian financial newspapers (i.e., the Australian Financial Review (AFR) and the National Business Review (NBR)). Results show that publishers consider hard (or fast-paced) news and opinion pieces as the most valuable news commodity. In addition, as presented, AFR locked 86% of its content compared to NBR's 41%.

In [10], authors analyze selected paywalled news sites in US, UK and Australia to compare the type, pricing and audience uptake. Results show that paywalls are part of newspapers' toolkit for bringing in new revenue but there is no evidence to suggest they can be a standalone solution. However, in this political economic environment for mastheads, digital advertising revenues alone are also insufficient to meet the cost of providing quality journalism.

8 CONCLUSION

Despite the seemingly important implications, paywalls impose on the free web, as an internet phenomena, they have been understudied. This paper aims to address this blind spot by conducting the first large scale study of paywalls on the web. Our results show that paywall use increases over time (2× more paywalls every 6 months), its adoption differs by country (e.g., 18.75% in US, 12.69% in Australia), and besides the privacy implications, paywalls fail to reliably protect publishers content. Finally, we present the design of a novel, automated system for detecting whether a site uses a paywall. We hope this work can be a significant first step in understanding the phenomena of paywalls.

REFERENCES

[1] Adam. 2018. Bypass Paywalls for Chrome. https://github.com/iamadamdev/bypass-paywalls-chrome.

[2] Andy Appleby. 2016. MurmurHash3. https://github.com/aappleby/smhasher/wiki/MurmurHash3.

[3] Ricardo Bilton. 2018. Learning from the New Yorker, Wired's new paywall aims to build a more "stable financial future". http://www.niemanlab.org/2018/02/learning-from-the-new-yorker-wireds-new-paywall-aims-to-build-a-more-stable-financial-future/.

[4] Christian Bonnie. 2017. Google's latest move means you actually have to pay for news. https://www.wired.co.uk/article/google-ditches-first-click-free-policy.

[5] Casey Oppenheim Brian Kennish. 2019. Disconnect Browser plugin. https://disconnect.me.

[6] Martin Brinkmann. 2016. Read articles behind paywalls by masquerading as Googlebot. https://www.ghacks.net/2016/02/26/read-articles-behind-paywalls-by-masquerading-as-googlebot/.

[7] Ryan Brown. 2019. Fanboy's Enhanced Tracking List. https://github.com/ryanbr/fanboy-adblock/blob/master/enhancedstats-addon.txt.

[8] Michael Burgi. 2016. What's Being Done to Rein In $7 Billion in Ad Fraud. https://www.adweek.com/brand-marketing/whats-being-done-rein-7-billion-ad-fraud-169743/.

[9] Ian Burrell. 2018. The FT will next year hit 1m subscribers, 17 years after putting up its paywall. https://www.thedrum.com/opinion/2018/08/30/the-ft-will-next-year-hit-1m-subscribers-17-years-after-putting-up-its-paywall.

[10] Andrea Carson. 2015. Behind the newspaper paywall–lessons in charging for online content: a comparative analysis of why Australian newspapers are stuck in the purgatorial space between digital and print. *Media, Culture & Society* 37, 7 (2015), 1022–1041.

[11] Nicholas Conley. 2018. The Problem with Paywalls. https://nicholasconley.wordpress.com/2018/05/01/the-problem-with-paywalls/.

[12] Alessio Cornia, Annika Sehl, Felix Simon, and Rasmus Kleis Nielsen. 2017. Pay models in European news. *Reuters Institute for the Study of Journalism, University of Oxford* (2017).

[13] Florent Daigniere. 2017. A browser extension that maximizes the chances of bypassing paywalls. https://github.com/nextgens/anti-paywall.

[14] Jerome Dangu. 2018. Uncovering 2017's Largest Malvertising Operation. https://blog.confiant.com/uncovering-2017s-largest-malvertising-operation-b84cd38d6b85.

[15] Jessica Davies. 2019. Ghost sites, domain spoofing, fake apps: A guide to knowing your ad fraud. https://digiday.com/media/ghost-sites-domain-spoofing-fake-apps-guide-knowing-ad-fraud/.

[16] Jeff Dunn. 2016. Wikipedia is asking for donations again — here's how much cash it already has in the bank. https://www.businessinsider.com/wikipedia-donations-profit-money-chart-2016-11.

[17] Rani Molla Edmund Lee. 2018. The New York Times digital paywall business is growing as fast as Facebook and faster than Google. https://www.recode.net/2018/2/8/16991090/new-york-times-digital-paywall-business-growing-fast-facebook-google-newspaper-subscription.

[18] eMarketer. 2017. Google and Facebook Tighten Grip on US Digital Ad Market. https://www.emarketer.com/Article/Google-Facebook-Tighten-Grip-on-US-Digital-Ad-Market/1016494.

[19] Michael Firn. 2018. Optimize Dynamic Paywall Conversions with Machine Learning. https://www.vidora.com/product-updates/dynamic-paywalls/.

[20] Mohammad Ghasemisharif, Peter Snyder, Andrius Aucinas, and Benjamin Livshits. 2018. SpeedReader: Reader Mode Made Fast and Private. *arXiv preprint arXiv:1811.03661* (2018).

[21] Roy Greenslade. 2013. Soft paywalls retain more users than hard paywalls - by a big margin. https://www.theguardian.com/media/greenslade/2014/nov/07/paywalls-charging-for-content.

[22] Internet Archive. 2001. Wayback Machine. https://archive.org/web/.

[23] Ho Kim, Reo Song, and Youngsoo Kim. 2019. Newspapers' Content Policy and the Effect of Paywalls on Pageviews. *Journal of Interactive Marketing* (2019).

[24] kufii. 2019. Newspaper Paywall Bypasser. https://greasyfork.org/en/scripts/18585-newspaper-paywall-bypasser/code.

[25] Richard Lawler. 2018. Apple seeks major newspaper allies for its subscription bundle. https://www.engadget.com/2018/09/08/apple-seeks-major-newspaper-allies-for-its-subscription-bundle/.

[26] Christophe Leung, Jingjing Ren, David Choffnes, and Christo Wilson. 2016. Should You Use the App for That?: Comparing the Privacy Implications of App- and Web-based Online Services. In *Proceedings of the 2016 Internet Measurement Conference (IMC '16)*. ACM, New York, NY, USA, 365–372. https://doi.org/10.1145/2987443.2987456

[27] Bin Liu, Suman Nath, Ramesh Govindan, and Jie Liu. 2014. DECAF: Detecting and Characterizing Ad Fraud in Mobile Apps. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation (NSDI'14)*. USENIX Association, Berkeley, CA, USA, 57–70. http://dl.acm.org/citation.cfm?id=2616448.2616455

[28] Mike Masnick. 2018. The Media's Paywall Obsession Will End In Disaster For Most. https://www.techdirt.com/articles/20180506/11501539779/medias-paywall-obsession-will-end-disaster-most.shtml.

[29] John McCarthy. 2019. The major trends shaping, breaking and consolidating global media by 2021. https://www.thedrum.com/news/2019/04/18/the-major-trends-shaping-breaking-and-consolidating-global-media-2021.

[30] Glyn Moody. 2013. Surprise: Paywalls Cause Massive Falls In Number Of Visitors - And Boost Competitors. https://www.techdirt.com/articles/20130920/09592024590/surprise-paywalls-cause-massive-falls-number-visitors-boost-competitors.shtml.

[31] Lucia Moses. 2017. The Wall Street Journal to close Google loophole entirely. https://digiday.com/media/wall-street-journal-close-google-loophole-entirely/.

[32] Mozilla Foundation. 2019. Readability.js. https://github.com/mozilla/readability.

[33] Merja Myllylahti. 2017. What Content is Worth Locking Behind a Paywall? Digital news commodification in leading Australasian financial newspapers. *Digital Journalism* 5, 4 (2017), 460–471.

[34] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J Murdoch. 2016. Adblocking and counter blocking: A slice of the arms race. In *6th {USENIX} Workshop on Free and Open Communications on the Internet ({FOCI} 16)*.

[35] Rodrigo Orem and Caio. 2018. Burlesco: Read news without subscribing, bypass the paywall. https://burles.co/en/.

[36] Outline. 2018. Outline - Read & annotate without distractions. https://outline.com/.

[37] Michalis Pachilakis, Panagiotis Papadopoulos, Evangelos P Markatos, and Nicolas Kourtellis. 2019. No More Chasing Waterfalls: A Measurement Study of the Header Bidding Ad-Ecosystem. In *Proceedings of the 19th Internet Measurement Conference (IMC 2019)*.

[38] Elias P Papadopoulos, Michalis Diamantaris, Panagiotis Papadopoulos, Thanasis Petsas, Sotiris Ioannidis, and Evangelos P Markatos. 2017. The long-standing privacy debate: Mobile websites vs mobile apps. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 153–162.

[39] Panagiotis Papadopoulos, Panagiotis Ilia, and Evangelos P Markatos. 2018. Truth in Web Mining: Measuring the Profitability and Cost of Cryptominers as a Web Monetization Model. *arXiv preprint arXiv:1806.01994* (2018).

[40] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos Markatos. 2019. Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*. 1432–1442.

[41] Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P Markatos. 2018. The cost of digital advertisement: Comparing user and advertiser views. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1479–1489.

[42] Panagiotis Papadopoulos, Nicolas Kourtellis, Pablo Rodriguez Rodriguez, and Nikolaos Laoutaris. 2017. If you are not paying for it, you are the product: How much do advertisers pay to reach you?. In *Proceedings of the 2017 Internet Measurement Conference*. ACM, 142–156.

[43] Ben Parr. 2011. What Impact Has The New York Times Paywall Had on Traffic? [STATS]. https://mashable.com/2011/04/11/new-york-times-paywall-stats/.

[44] Piano Inc. 2015. Overview - Tinypass for Developers. http://developer.tinypass.com/.

[45] Piano Inc. 2019. Algorithmic Paywall. https://docs.piano.io/algorithmic-paywall/.

[46] Dominic Ponsford. 2017. Press Gazette launches Duopoly campaign to stop Google and Facebook destroying journalism. https://www.pressgazette.co.uk/press-gazette-launches-duopoly-campaign-to-stop-google-and-facebook-destroying-journalism/.

[47] publicWWW. 2019. Source Code Search Engine. https://publicwww.com/.

[48] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, and Christian Kreibich Phillipa Gill. 2018. Apps, trackers, privacy, and regulators. In *Proceedings of the Network and Distributed System Security Symposium (NDSS'18)*.

[49] Patricio Robles. 2017. Google ditches first click free, embraces paywalls. https://econsultancy.com/google-ditches-first-click-free-embraces-paywalls/.

[50] Janko Roettgers. 2017. BuzzFeed CEO Jonah Peretti: Paywalls Are Bad for Democracy. https://variety.com/2017/digital/news/buzzfeed-jonah-peretti-paywalls-democracy-1202593489/.

[51] Sam Rutherford. 2018. Google Thinks It Can Make Paywalls Less of a Pain in the Ass. https://gizmodo.com/google-thinks-it-can-make-paywalls-less-of-a-pain-in-th-1823925612.

[52] Felix Salmon. 2011. The NYT paywall is working. http://blogs.reuters.com/felix-salmon/2011/07/26/the-nyt-paywall-is-working/.

[53] Zach Saucier. 2018. Just Read. https://justread.link/.

[54] Scikit-learn developers. 2019. Scikit-learn: Machine Learning in Python. https://scikit-learn.org/stable/index.html.

[55] Felix Simon and Lucas Graves. 2019. Across seven countries, the average price for paywalled news is about $15.75/month. https:

//www.niemanlab.org/2019/05/across-seven-countries-the-average-price-for-paywalled-news-is-about-15-75-month/.

[56] Lucinda Southern. 2018. How Swiss news publisher NZZ built a flexible paywall using machine learning. https://digiday.com/media/swiss-news-publisher-nzz-built-flexible-paywall-using-machine-learning/.

[57] Duncan Stewart. 2018. Are Consumers 'Adlergic'? A Look at Ad-Blocking Habits. https://deloitte.wsj.com/cmo/2018/04/03/are-consumers-adlergic-a-look-at-ad-blocking-habits/.

[58] Fitz Tepper. 2017. Facebook is now testing paywalls and subscriptions for Instant Articles. https://techcrunch.com/2017/10/19/facebook-is-now-testing-paywalls-and-subscriptions-for-instant-article/.

[59] Narseo Vallina-Rodriguez, Srikanth Sundaresan, Abbas Razaghpanah, Rishab Nithyanand, Mark Allman, Christian Kreibich, and Phillipa Gill. 2016. Tracking the trackers: Towards understanding the mobile advertising and tracking ecosystem. *arXiv preprint arXiv:1609.07190* (2016).

[60] Antoine Vastel, Peter Snyder, and Benjamin Livshits. 2018. Who filters the filters: Understanding the growth, usefulness and efficiency of crowdsourced ad blocking. *arXiv preprint arXiv:1810.09160* (2018).

[61] Shan Wang. 2018. After years of testing, The Wall Street Journal has built a paywall that bends to the individual reader. http://www.niemanlab.org/2018/02/after-years-of-testing-the-wall-street-journal-has-built-a-paywall-that-bends-to-the-individual-reader/.

[62] Nate Weiner. 2007. Pocket - Put knowledge in your Pocket. https://getpocket.com.

[63] Wikimedia Foundation. 2018. 2016-2017 Fundraising Report. https://foundation.wikimedia.org/wiki/2016-2017$_{\text{F}}$undraising$_{\text{R}}$eport.

[64] Apostolis Zarras, Alexandros Kapravelos, Gianluca Stringhini, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. 2014. The Dark Alleys of Madison Avenue: Understanding Malicious Advertisements. In *Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14)*.